

# 2019/4/15 - 2019/4/21 Weekly Report

Sunday, April 21, 2019

2:16 PM

王智勇

## 回顾

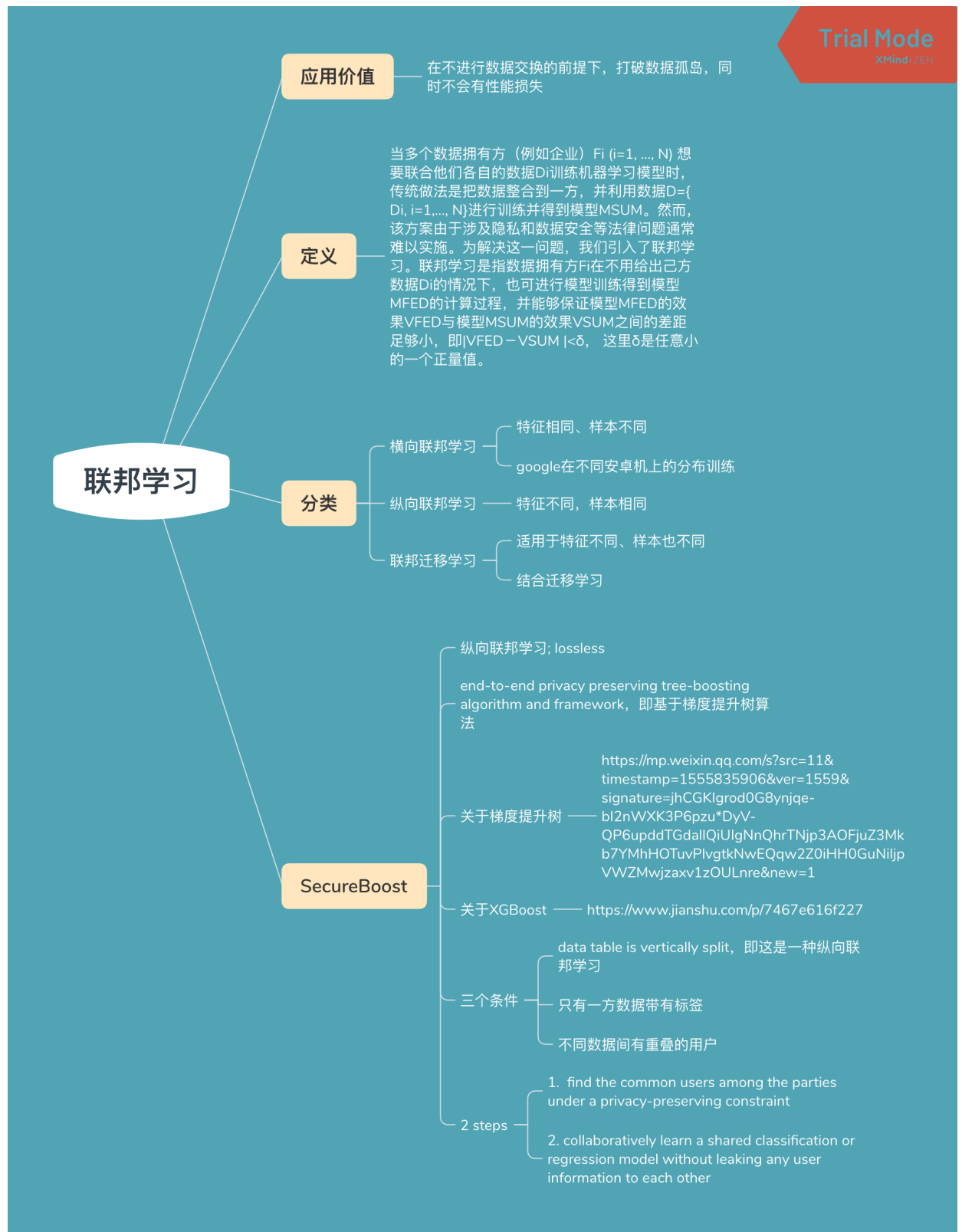
### 1. 概况

- 实验室项目方面：学习联邦学习，为小组项目做准备；
- 自主学习方面：
- 学校事宜：学校课程小组项目。

### 2. 时间安排

时间	主要内容	实验室工作时长
周一	实验室	8h
周二	实验室	8h
周三	学校课程项目	0
周四	学校课程项目	0
周五	学校课程项目	0
周六	学习联邦学习	8h
周日	休息	0

### 3. 联邦学习



#### 学习途径

#### - 阅读材料：

- [xumeng周报](#)：
- [GDPR对AI的挑战和基于联邦迁移学习的对策\[J\]](#)

- [王晋东-联邦迁移学习](#)
- [Federated Machine Learning: Concept and Applications](#)
- [SecureBoost: A Lossless Federated Learning Framework](#)

## 应用价值

不进行数据交换的情况下打破数据孤岛，且性能等同于数据聚合的情况。

- 示例：

假设有两个不同的企业A和B，它们拥有不同的数据，比如企业A有用户特征数据，企业B有产品特征数据和标注数据。这两个企业按照GDPR准则是不能粗暴地把双方数据加以合并的，因为他们各自的用户并没有机会同意这样做。假设双方各自建立一个任务模型，每个任务可以是分类或预测，这些任务也已经在获得数据时取得了各自用户的认可。那么，现在的问题是如何在A和B各端建立高质量的模型。但是，又由于数据不完整（例如企业A缺少标签数据，企业B缺少特征数据），或者数据不充分（数据量不足以建立好的模型），各端有可能无法建立模型或效果不理想。联邦学习的目的是解决这个问题：它希望做到各个企业的自有数据不出本地，联邦系统可以通过加密机制下的参数交换方式，在不违反数据隐私保护法规的情况下，建立一个虚拟的共有模型。这个虚拟模型就好像大家把数据聚合在一起建立的最优模型一样。但是在建立虚拟模型的时候，数据本身不移动，也不会泄露用户隐私或影响数据规范。这样，建好的模型在各自的区域仅为本地的目标服务。在这样一个联邦机制下，各个参与者的身份和地位相同，而联邦系统帮助大家建立了“共同富裕”的策略。这就是为什么这个体系叫做“联邦学习”。

来自 <<https://dl.ccf.org.cn/institute/instituteDetail?id=4150944238307328>>

- 微众银行应用场景

联邦学习的提出，是为了应对金融机构的痛点，尤其是像“微众银

行”这样的互联网银行。其中一个实例是检测多方借贷。这在银行业，尤其是互联网金融行业一直是一个很头疼的问题。多方借贷是指某不良用户在一个金融机构借贷后还钱给另一个借贷机构，大量这种非法行为会让整个金融系统崩溃。要想发现这样的用户，传统的做法是金融机构去某中心数据库查询用户信息，而各个机构必须上传他们所有的用户信息，但这样做等于暴露了金融机构的所有重要用户隐私和数据安全，这在GDPR下是不允许的。在联邦学习机制下，没有必要建立一个中心数据库，而任何参与联邦学习的金融机构可以向联邦内的其他机构发出新用户查询请求，其他机构在不知道这个用户具体信息的前提下，回答该用户关于本地借贷的提问。这样既能保护已有用户在各个金融机构的隐私和数据完整性，同时也能完成查询多方借贷这个重要问题。

来自 <<https://dl.ccf.org.cn/institute/instituteDetail?id=4150944238307328>>

- 联邦学习系统

让各个机构或公司组织加入，借助联邦学习，是的各方彼此不交换数据，也能利用彼此的数据更好地训练自己的模型。另外，为了构建可信的联邦迁移学习系统，各个企业应当在遵循法律法规的基础上，按照各参与方理解一致的共识机制，构建基于区块链的运营组织。区块链使得信息的存储变得去中心化，从而避免了信息泄露和伪造。

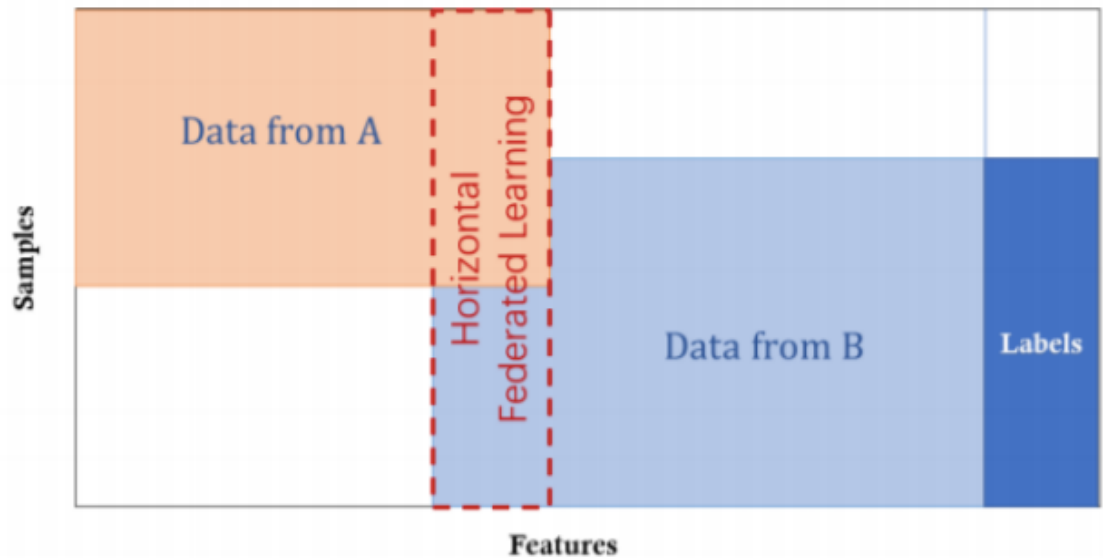
## 定义

当多个数据拥有方（例如企业） $F_i$  ( $i=1, \dots, N$ ) 想要联合他们各自的数据 $D_i$ 训练机器学习模型时，传统做法是把数据整合到一方，并利用数据 $D=\{D_i, i=1, \dots, N\}$ 进行训练并得到模型 $M_{SUM}$ 。然而，该方案由于涉及隐私和数据安全等法律问题通常难以实施。为解决这一问题，我们引入了**联邦学习**。联邦学习是指数据拥有方 $F_i$ 在不用给出己方数据 $D_i$ 的情况下，也可进行模型训练得到模型 $M_{FED}$ 的计算过程，并能够保证模型 $M_{FED}$ 的效果 $V_{FED}$ 与模型 $M_{SUM}$ 的效果 $V_{SUM}$ 之间的差距足够小，即 $|V_{FED} - V_{SUM}| < \delta$ ，这里 $\delta$ 是任意小的一个正量

值。

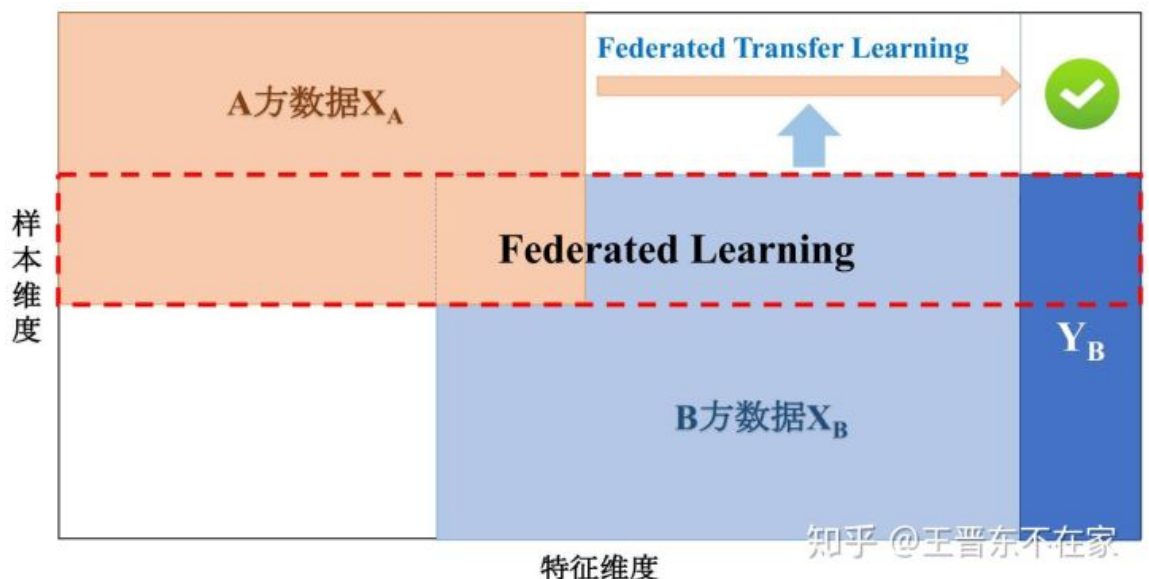
## 分类

### 1) 横向联邦学习



来自不同机构的的数据具有不同的样本，有着相同的特征。如中国移动和中国联通的数据来自于不同的用户，但具有类似的维度特征。最早提出联邦学习的Google也是利用这种方式，在众多手机上的进行联邦学习，而无需让用户数据上传到服务器。

### 2) 纵向联邦学习

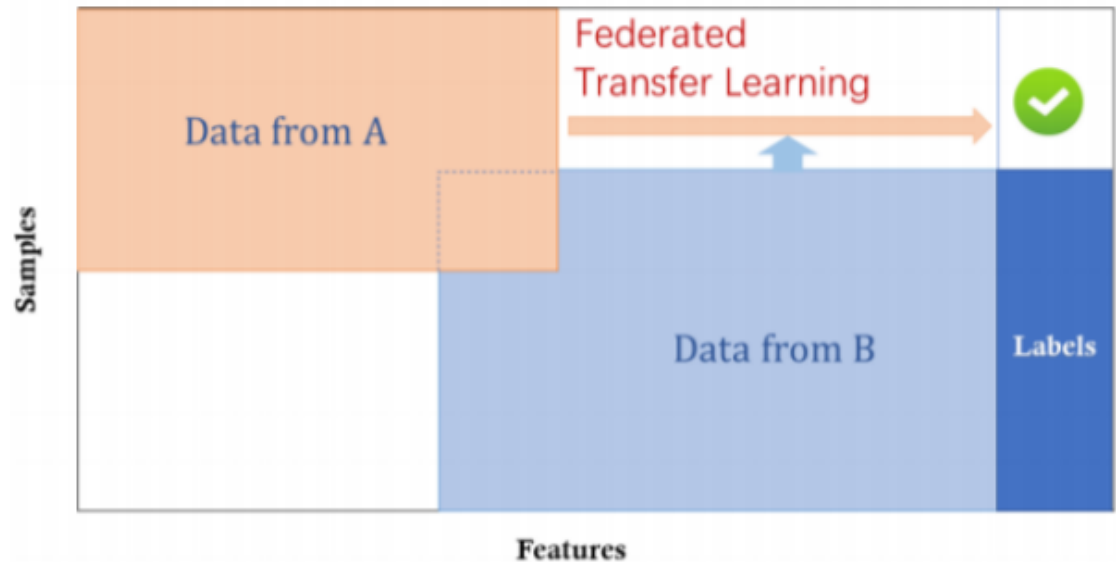


来自不同机构的数据有着共同的一批样本，但具有不同的特征。如淘宝和联通，他们可能具有同一批用户（相同的手机号）的数据，但数据特征不同，这时候使用联邦学习的效果相当于将两个公司的

数据joint之后进行训练的。

SecureBoost就是一种纵向联邦学习：[SecureBoost: A Lossless Federated Learning Framework](#)

### 3) 联邦迁移学习



结合迁移学习，让有着不同样本和不同维度的数据也能结合利用。

阅读：[Secure Federated Transfer Learning](#)

### 联邦迁移学习vs迁移学习vs多任务学习

多任务学习和联邦迁移学习都注重多个任务的协同学习，最终目标都是要把所有的模型变得更强。但是，多任务学习强调不同任务之间可以共享训练数据，破坏了隐私规则；而联邦迁移学习则可以在不共享隐私数据的情况下，进行协同的训练。

迁移学习注重知识从一个源领域到另一个目标领域的单向迁移。而这种单向的知识迁移，往往伴有一定的信息损失：因为我们通常只会关注迁移学习在目标领域上的效果，而忽略了在源领域上的效果。联邦迁移学习则从目标上就很好地考虑了这一点：多个任务之间协同。

### 浅入SecureBoost

- 文献：[SecureBoost: A Lossless Federated Learning Framework](#)
- SecureBoost是基于梯度提升树算法一种纵向联邦学习（end-to-end privacy preserving tree-boosting algorithm and framework）

framework)。

- 关于梯度提升树与XGBoost

[梯度提升树介绍](#)

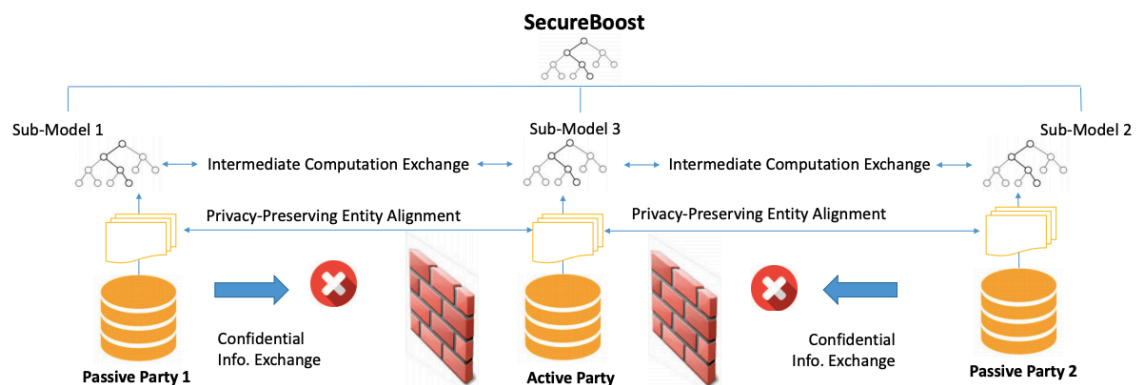
[XGBoost介绍](#) [文献](#)

- 该算法的三个条件

data table is vertically split , 即这是一种纵向联邦学习  
只有一方数据带有标签  
不同数据间有重叠的用户

- 2 steps

1. find the common users among the parties under a privacy-preserving constraint
2. collaboratively learn a shared classification or regression model without leaking any user information to each other



Active Party表示带有标签的数据，Passive Party表示无标签的数据

( 该部分有待补充 )

## 计划

### 1. 下周

- 继续学习联邦学习，细节到具体算法；了解相关金融背景；为之后微众银行项目做准备；

### 2. 近期

- 系统学习可视化领域知识。目标：理解当前可视化领域中热门的研究方向。

- 学习GNN领域内容。目标：能够从原理出发理解GNN的应用领域及优势；能够在将算法应用在某些数据上。
- 熟悉联邦学习及相关金融领域知识，参与微众银行项目。